EURASIAN BULLETIN OF MATHEMATICS
EBM (2021), VOL. 4, NO. 1, 32-37

Estimating a Finite Population Total using a Density Function

Dioggban Jakperik[1,*], Romanus Otieno Odhiambo[2], and Jacob Okungu[2]

Department of Statistics, Faculty of Mathematical Sciences, CK Tedam University of Technology and Applied Sciences, Box 24, Navrongo, Ghana[1]
Department of Mathematics, School of Pure and Applied Sciences, Meru University of Science and Technology, P. O. Box 972-60200, Meru, Kenya[2]

Abstract. An improved method for finite population total estimation is proposed using a multiplicative semi-parametric bias reduction density function. The density is first applied to estimate a non-parametric regression model which describes the relationship between the study variable and the auxiliary variable. For each value of the study variable, there exist a corresponding value of the auxiliary variable in the population. The proposed estimator is compared to the expansion estimator and the Nadarya-Watson estimator using bandwidths ($h = 0.25, 0.5, 0.75$) respectively through a simulation study using the Ghana Living Standards Survey Round Six data. The proposed estimator performed better than its competitors yielding the lowest Root Average Squared Error (ARE). The estimator can be applied to datasets with high variances without any transformations. its optimum efficiency and precision are achieved when the sample size is large.
2010 Mathematics Subject Classifications: 16P40; 13A15; 16D60
Key Words and Phrases: Non parametric regression, Multiplicative semi-parametric bias reduction density function, Relative Bias, Root Average Squared Error

1. Introduction

Sample survey has become very important in our daily life due to the alternatives it offers in undertaking feasible policy decisions at a faster rate and relatively cheaper cost compared to census among others. This therefore preoccupied survey statisticians to finding credible methods of improving precision of statistics emanating from surveys; using kernel and local linear regressions [9], kernel regression with transformations [5], and improved Nadaraya-Watson estimator [1]. Hence statistics such as population total, mean, and variance are often perceived to provide satisfactory estimates of their population counterparts. Studies have rather sought to agree that in-sample and out-sample observations (values) are generated from the same underlying distribution but not necessarily the same and hence estimates of this underlying distribution should be obtained and used to predict the out-sample observations for improved precision [8, 10, 2].

These studies, mostly non parametric in nature have achieved considerable success but still needs an improvement. For instance boundary problems especially in situations where sample

*Corresponding author.

Email address: jdioggban@cktutas.edu.gh* (Corresponding Author)

auxilliary variables are not spread throughout the non-sample values, the kernel estimation process is expected to experience challenges [5]. In this study, a multiplicative semi-parametric bias reduction density function is used to develop a non parametric estimator for a finite population total.

The outline of the paper is as follows: in section 2 an overview of non-parametric regression is discussed, the multiplicative semi-parametric bias reduction density which is used to derive the proposed estimator is given in section 3, the various methods used in non-parametric regression modeling are stated in section 4, the proposed estimator of the population total is given in section 5, simulation study is conducted in section 6 with conclusion in section 7.

2. Overview of non Parametric Regression

The concept of non-parametric regression was proposed by [8] and [10], it has a general form

$$Y = m(x) + e$$

where $m(.)$ is a smooth function and e_i is an independent error with mean 0 and constant variance. In estimating $m(.)$, a model is often assumed between the variable of interest and the auxilliary variable. Wrong specification of the model could lead to efficiency loss. Therefore, the ultimate approach is to use the Nadaraya-Watson estimator which uses the kernel density estimator. In its implementation, it computes the averages of data points which falls within the bandwidth, h around the knots. The weight is therefore given by

$$w_i(x) = K_h(x_i - x) / \sum_{i=1}^{n} K_h(x_i - x)$$

The Nadaraya-Watson estimate of $m(.)$ is thus given by

$$\hat{m}(x) = \sum_i w_i(x) Y_i$$

Under regularity conditions, $\hat{m}(x)$ is consistent for $m(x)$ as $h \to 0$ and $nh \to \infty$. Also, $\hat{m}(x)$ is enhanced for $h \to 0$ since large values of h leads to the density function estimate to be flatter and broad, hence increases estimation bias.

3. Multiplicative Semi-parametric Bias Reduction Density Function

In this study, a multiplicative semi-parametric bias reduction density estimator is used in the estimation of the non-parametric smooth equation. The general form of the density is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x) \frac{f\left(x, \hat{\theta}\right)}{f\left(X_i, \hat{\theta}\right)}$$

where $K_h(X_i - x)$ is the kernel density, n is the sample size, $f\left(x, \hat{\theta}\right)$, and $f\left(X_i, \hat{\theta}\right)$ are non-parametric and parametric density estimators respectively.

The multiplicative semi-parametric bias reduction density estimator has a bias

$$bias\left(\hat{f}(x)\right) = \frac{1}{24} h^4 \mu_4 f_0 r^{(iv)}(x) + o\left(h^4\right)$$

and variance given by

$$var\left(\hat{f}(x)\right) = \frac{R(K)}{nh} f(x) - \frac{f(x)^2}{n}$$

Details and properties of this estimator can be found in [7].

4. Estimation of Population Total Based on the Non-Parametric Regression

Consider a population U of size N units for which there exists a variable Y from a sample s of the population U. Suppose x_j represents the out-sample observations, the estimate of $m(x_j)$ is obtained. The estimator of the population total is given by

$$\hat{T}_{sp} = \sum_s Y_i + \sum_{p-s} \hat{m}(x_j)$$

This estimator does not use sampling probabilities and stratum boundaries. It is an automatic estimator except for bandwidth selection and any possible transformation of auxilliary variable. The population total is defined by $T_U = \sum_U Y_i$ whilst estimate of the population total based on a sample s is $\hat{T}_s = \sum_s Y_i$. Assuming for each Y_i, there is also an auxilliary variable x which closely relates to the Y_i's, then a non parametric regression of Y on x could provide predicted values of Y's to enhance estimate and/or precision of the population total.

In model based estimation of the population total, the common source(s) of bias has always been wrong specification of the model between the auxilliary variable and the variable of interest, wrong specification of the model form, especially as most researchers adopt the linear model for its tractability which in most cases do not achieve much efficiency gain over purely design-based estimators. This therefore called for consideration of non-parametric modeling to enhance estimation efficiency and precision. Various approaches to non-parametric modeling as regards the density functions have been proposed: kernel density using Nadaraya-Watson estimator [5, 4] and model-assisted estimators based on local polynomial smoothing [3, 9]. The local polynomial regression estimator has the form of the generalized regression estimator, but is based on a non-parametric superpopulation model applicable to a much larger class of functions. Its automaticity has helped in increasing precision of parameter estimates and hence has widely been used in recent times.

5. Proposed Estimator of Population Total

Let $x = x_j$ for any non-sample observation and so estimate $m(x_j)$. Then the following estimator of the total suggests itself

$$\hat{T}_{sp} = \sum_s Y_i + \sum_{p-s} \hat{m}(x_j) \tag{5.1}$$

where

$$\hat{m}(x_j) = \sum_j w_j(x) Y_j$$

and using

$$w_j(x) = \frac{\hat{f}(x)}{\sum_{j=1}^n \hat{f}(x)} \tag{5.2}$$

Using the substitution, $z = \frac{x-x_j}{h}$ and $f(x) := f(x+hz)$. Applying Tailor series expansion gives,

$$f(x+hz) = f(x) + o\left(h^2\right)$$

substituting into (5.2) together with (5.1) gives

$$\begin{aligned}
\hat{T}_{sp} &= \sum_s Y_i + \sum_{P-s}\left(Y_j / \sum_{j=1}^{n} \hat{f}(x)\right)\left[f(x) + o\left(h^2\right)\right] \\
&= \sum_s Y_i + \sum_{P-s}\left(Y_j / \sum_{j=1}^{n} \hat{f}(x)\right) f(x) + \sum_{P-s}\left(Y_j / \sum_{j=1}^{n} \hat{f}(x)\right) o\left(h^2\right)
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow E\left(\hat{T}_{sp}\right) &= \sum_s Y_i + \sum_{P-s}\left(Y_j / \sum_{j=1}^{n} \hat{f}(x)\right) f(x) \\
&= \sum_s Y_i + \sum_{P-s}\left(Y_j / \sum_{j=1}^{n} \hat{f}(x)\right) f(x)
\end{aligned}$$

Consequently, the bias of the estimator is given by

$$\hat{T}_{sp} - E\left(\hat{T}_{sp}\right) = o\left(h^2\right)$$

For stratified population

$$\begin{aligned}
\hat{T}_{sp} &= \sum_s Y_i + \sum_{P-s} \hat{m}_j(x) \\
&= \sum_{h=1}^{l} \frac{N_h - n_h}{N_h}\left(\sum_s Y_i + \sum_{P-s} \hat{m}_j(x)\right) \\
&= \sum_{h=1}^{l} \frac{N_h - n_h}{N_h} \sum_s Y_i + \sum_{h=1}^{l} \frac{N_h - n_h}{N_h} \sum_{P-s} \frac{f(x)}{\sum f(x)} Y_j \\
&= \sum_{h=1}^{l} \frac{N_h - n_h}{N_h}\left[\sum_s Y_i + \sum_{P-s} \frac{f(x)}{\sum f(x)} Y_j\right]
\end{aligned}$$

6. Simulations

The simulation study was based on data from the Ghana Living Standards Survey Round 6 [6], with the ten administrative regions representing the strata especially for application of the expansion estimator. It used a sample of size 1000 drawn from the population and fitted to non-parametric regression for bandwidths $h = 0.25, 0.5, 0.75$ and the semi-parametric estimators over 1000 iterations. In each case, the accuracy measures such as the average relative error, $ARB = \sum_{r=1}^{100} T^{-1}\left(\hat{T}_r - T\right)/1000$ and the root average squared error, $RSE = \sum_{r=1}^{1000}\left(\hat{T}_r - T\right)^2/1000$, where $\hat{T}_r$ is one of the estimators of T computed for sample r. The results of these simulations are given in table 1.

Table 1: **ACCURACY MEASURES FOR THE ESTIMATORS**

Estimator	**ARE**	**RSE**$\times 10^{13}$	**RASE** $\left(\hat{\mathbf{T}}_{\mathbf{R}}\right)$/**RASE** $\left(\hat{\mathbf{T}}_{\mathbf{exp}}\right)$
Semi-parametric	-0.3872	1.1345	1.2142
Expansion	5.3659	1.3775	1
Non-parametric reg ($h = 0.25$)	0.2796	5.6201	0.2451
Non-parametric reg ($h = 0.5$)	0.2005	4.0338	0.3415
Non-parametric reg ($h = 0.75$)	0.1043	2.5758	0.5348

The simulation results revealed an interesting phenomenon with respect to the performance of the various estimators. Whilst the RSE of the Expansion estimator is only next to the Semi-parametric estimator, it's ARB is the highest. This could probably be due to the extremes in the simulation data. The Semi-parametric estimator on the other hand, emerged as the best estimator having produced the smallest RSE against all the estimators, yielding at least 21% efficiency more than the expansion estimator which is a model-based estimator with comparable efficiency as a design-based estimator. This may likely be due to the fact that the Semi-parametric estimator is a large sample estimator and also has an automatic correction for boundary bias. The semi-parametric estimator however had higher ARB than all the regression estimators. The efficiency of the non-parametric regression estimator increased with increasing bandwidths which seeks to confirm the notion that the underlying data used for the study had much variability and thus increasing bandwidth increases the number of data points used for estimation and hence estimation precision. However, the best performance of the regression estimator is achieved at $h = 0.75$ which is 53.48%. This is almost half the efficiency of the expansion estimator. This probably could be attributed to the presence of outliers and extreme variability. Because the regression estimator does not automatically correct for boundary bias, it negatively impacted its performance hence reduced its efficiency.

7. Conclusion

A semi-parametric multiplicative bias reduction density function has been used to develop an estimator for the estimation of the finite population total. Simulation studies conducted justified its practical potential as it performed better than all the estimators that were compared to it.

References

[1] Bii, N. K., Onyango, C. O. and Odhiambo, J. (2020). Estimation of a Finite Population Mean under Random Nonresponse Using Kernel Weights. Journal of Probability and Statistics(Hindawi).

[2] Bii, N. K. and Onyango, Ouma C. (2017). Model-Assisted Estimation of Population Mean in Two-Stage Cluster Sampling. Pakistan Journal of Statistics and Operation Research, 13(1):127–139.

[3] Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. Annals of Statistics.

[4] Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. Journal of the American Statistical Association, 88(421): 268–277.

[5] Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. Proceedings of the Section on Survey Research Methods, Citeseer.

[6] Ghana Statistical Service (2014). Ghana Living Standards Survey Round 6 (GLSS 6).: Poverty profile in Ghana (2005-2013).

[7] Jakperik, D., Odhiambo, R. O. and Orwa, G. O. (2015). A semi-Parametric Multiplicative Bias Reduction Density with a Parametric Start. Advances in Statistics and Applications, 53(6):715-729.

[8] Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. Journal of Theory of Probability & Its Applications, 10(1): 186–190.

[9] Rady, E. H. A. and Ziedan, D. (2014). A new technique for estimation of total using non-parametric regression under two stage sampling. Journal of Applied Mathematical Sciences, 8(74):3647–3659.

[10] Watson, G. S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A: 359–372.